D396–D400 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 11 November 2009*
doi:10.1093/nar/gkp919

MicrobesOnline: an integrated portal for comparative and functional genomics

Paramvir S. Dehal[1,2,*], Marcin P. Joachimiak[1,2], Morgan N. Price[1,2], John T. Bates[1,2,3], Jason K. Baumohl[1,2], Dylan Chivian[1,2,3], Greg D. Friedland[1,2,3], Katherine H. Huang[1,2], Keith Keller[1,2], Pavel S. Novichkov[1,2], Inna L. Dubchak[1,2], Eric J. Alm[1,4] and Adam P. Arkin[1,2,3,5]

[1]Virtual Institute for Microbial Stress and Survival, [2]Lawrence Berkeley National Laboratory, Berkeley, CA 94720, [3]DOE Joint BioEnergy Institute, [4]Department of Biological Engineering and Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, [5]Department of Biological Engineering and [6]Department of Bioengineering, University of California, Berkeley, CA, 94720, USA

Received September 18, 2009; Accepted October 7, 2009

ABSTRACT

Since 2003, MicrobesOnline (http://www.microbesonline.org) has been providing a community resource for comparative and functional genome analysis. The portal includes over 1000 complete genomes of bacteria, archaea and fungi and thousands of expression microarrays from diverse organisms ranging from model organisms such as *Escherichia coli* and *Saccharomyces cerevisiae* to environmental microbes such as *Desulfovibrio vulgaris* and *Shewanella oneidensis*. To assist in annotating genes and in reconstructing their evolutionary history, MicrobesOnline includes a comparative genome browser based on phylogenetic trees for every gene family as well as a species tree. To identify co-regulated genes, MicrobesOnline can search for genes based on their expression profile, and provides tools for identifying regulatory motifs and seeing if they are conserved. MicrobesOnline also includes fast phylogenetic profile searches, comparative views of metabolic pathways, operon predictions, a workbench for sequence analysis and integration with RegTransBase and other microbial genome resources. The next update of MicrobesOnline will contain significant new functionality, including comparative analysis of metagenomic sequence data. Programmatic access to the database, along with source code and documentation, is available at http://microbesonline.org/programmers.html.

INTRODUCTION

MicrobesOnline seeks to integrate functional genomic data with comparative genome analysis (1) providing two unique capabilities: (i) a phylogenetic approach to comparative genomics, including a tree-based browser and tools for users to build their own tress, and (ii) microarray expression data integration. Selecting an organism or gene of interest in MicrobesOnline leads to information about and data viewers for experiments conducted on that organism and involving that gene or gene product. It is possible to view microarray data from multiple conditions as an interactive heatmap and to analyze correlations between gene expression results from different experiments. Among the major new features is the ability to search the microarray data compendium for genes with gene expression profiles similar to a query expression profile (either based on a gene or set of genes). These new compendium-wide functionalities allow the user to observe patterns in gene expression changes across multiple conditions and genes, and to search for similarities to these patterns. The information integration and analysis performed by MicrobesOnline serves not only to generate insights into the gene expression responses and their regulation in these microorganisms, but also to document experiments, allow contextual access to experimental data and facilitate the planning of future experiments. MicrobesOnline is actively

*To whom correspondence should be addressed. Tel: +1 510 643 3722; Email: psdehal@lbl.gov

The authors wish it to be known that, in their opinion the first and the ninth author should be regarded as joint First Authors.

incorporating publicly available functional genomics data from published research, so as to centralize data on microbial physiology and ecology in a unified comparative functional genomic framework.

DATA SOURCES AND ANALYSIS PIPELINE

Sequences and annotations of bacterial and archaea genomes, plasmids and viruses are imported from the NCBI RefSeq database (2). Additionally, sequences not yet deposited in RefSeq are also loaded into MicrobesOnline by request of users. All sequences are associated with their taxonomic lineage through mapping to the NCBI taxonomy database (3). Across all sequences (chromosomes, scaffolds, contigs, genes, proteins and RNAs), identifiers are maintained in order to link to other NCBI resources, such as PubMed, and to facilitate cross-links to other resources, such as IMG (4) and RegTransbase (5). For user-provided genomes and for incomplete genomes downloaded from a genome center, gene predictions are taken from the user or predicted by combining results from CRITICA (6) and GLIMMER3 (7); in either case, these predictions are superseded by the RefSeq predictions when they become available. CRISPR and RNA loci potentially missing from the RefSeq annotations are computed using PILER-CR (8), CRT (9), tRNAScan (10) and BLASTn.

Functional annotations of protein-coding genes include gene name, gene family, domain architecture, Enzyme Commission (EC) number (11), Gene Ontology (GO) terms (12) and user annotations. The gene name is provided by RefSeq, but all other annotations are inferred by MicrobesOnline using homology searches. FastBLAST and FastHMM (13) are used to search against the TIGRFAM (14), Superfamily (15), SMART (16), Pfam (17), Panther (18), PirSF (19), Gene3D (20), COG (21), PDB (22) and UniProt (23) databases. Characterized genes—genes that UniProt or RegTransBase link to papers (except genome papers)—are automatically highlighted on many MicrobesOnline pages. EC is from KEGG, GO from InterPro. Genes that link to UniProt papers (but not genome papers) or to RegTransBase papers are considered characterized, and are highlighted in many of our pages.

Using FastBLAST, we are also able to calculate all-against-all homology relationships for the entire microbial proteome. These results allow users to quickly find a list of homologs for any gene. Additionally, these shared regions of homology and each conserved domain alignment are used to create phylogenetic trees using the FastTree program.

Within each gene tree, MicrobesOnline identifies clades that are (mostly) present just once per genome. These are likely to be functional orthologs, but not necessarily evolutionary orthologs. MicrobesOnline uses these single-copy clades to compute MicrobesOnline ortholog groups, which are groups of likely functionally orthologous genes that are stored in the database. MicrobesOnline also computes higher quality 'tree orthologs' for a given query gene on demand. For details, see http://www.microbesonline.org/ortholog.html.

Gene expression data has been collected from primarily three sources, the DOE VIMSS, M3D (24) and NCBI GEO (25) database. Additionally, individual users can submit microarray data in order to use the MicrobesOnline analysis tools. MicrobesOnline contains only a partial import of the GEO database, but this will be expanded in future updates to MicrobesOnline.

BULK DOWNLOADS AND REMOTE ACCESS

We provide various methods for programmatic access to our data and code. For raw database access, we provide a public MySQL server that contains a version of MicrobesOnline with all private genomes removed. This database can be accessed using any MySQL compliant client or programmatic access through the standard MySQL libraries. Information on how to use the MySQL database, a detailed database schema description, the BLAST and FastBLAST databases and Perl libraries and source code can be found on the MicrobesOnline for Programmers page at http://www.microbesonline.org/programmers.html. Instructions for making contributions to the MicrobesOnline code base can also be found on the MicrobesOnline for Programmers page.

UPDATE AND DATA SUBMISSION PROCEDURES

MicrobesOnline has the capacity to host genomes and microarray expression data submitted by users. These data can be kept private through a user-defined access control list until publication or the submitter wishes to release the data publicly. MicrobesOnline accepts

Table 1. MicrobesOnline feature list

Comparative genomics
TreeBrowser
Phylogenetic profiler
Domains
Gene families
Orthologs
Protein structures
Operons
RNA genes
Known regulatory sites
CRISPRs
Functional genomics
GO
EC
KEGG pathway maps
Regulons
PubMed links to characterized genes
Gene carts
Multiple sequence alignments
Phylogenetic trees
Regulatory motifs
Microarrays
Co-expression correlations
Expression profile search

genomic data in any state whether finished sequence or draft sequence and annotated sequence or unannotated. Instructions for preparing the data can be found on at http://www.microbesonline.org/privateGenomeHosting.html. MicrobesOnline can also host private microarray data; however, this must be done by contacting help@microbesonline.org for detailed instructions.

MicrobesOnline FEATURE HIGHLIGHTS

MicrobesOnline provides a comprehensive set of interfaces and tools for comparative and functional genomics (Table 1). These tools perform operations on combinations of genomes, genes, predicted gene functions, gene expression data and metabolites. A detailed tutorial on the use of these tools is provided on the MicrobesOnline website (http://www.MicrobesOnline.org). Below we highlight three of the more novel aspects of the MicrobesOnline website: a phylogenetic tree based genome browser, a microarray expression profile search tool and a comparative browser for metabolic compound and pathways.

PHYLOGENETIC TreeBrowser

The phylogenetic TreeBrowser allows users to view a selected gene within its evolutionary context (Figure 1). Viewing genes within a phylogenetic tree is not only a more biologically meaningful, but concisely summarizes the thousands of pair-wise alignments from all against all BLAST results into a single hypothesis of homology. The TreeBrowser has three viewing modes, a gene context view, a species tree view and a domains view. In the gene context viewing mode, for any selected gene, the tree browser displays a phylogenetic tree for a selected gene along with genomic context of that gene and its close relatives in the gene tree. These trees are based on a user-selectable multiple sequence alignment that ranges from the full gene sequence to a known domain families or *ad hoc* BLAST-based clusters. The genome context, the neighboring upstream and downstream genes are shown colored by the gene family. This allows the user to quickly scan the tree to determine conserved genomic context. Because of the relatively large size of microbial gene families, which can be over 100 000 for the largest families, it is necessary to collapse branches and show only a single representative for each clade. The

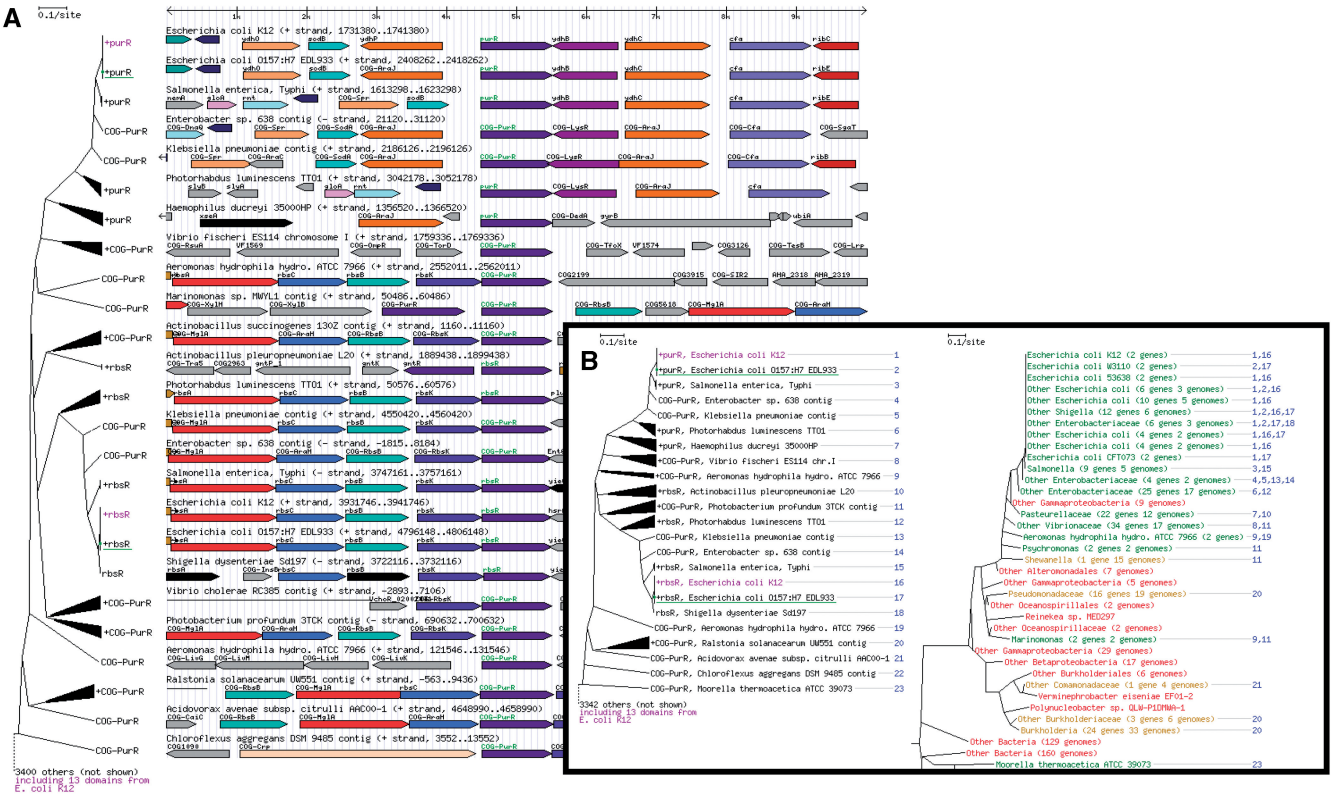

Figure 1. TreeBrowser in gene centric view and species view. (**A**) The TreeBrowser gene centric view for the *E. coli* K12 gene purR. The purR gene is shown centered in the browser with the neighboring up- and downstream genes. The phylogenetic tree for the purR gene is shown on the left, with a single representative selected for each clade. (**B**) The species view for the purR gene. The gene tree is shown on the left and the species tree on the right. Gene tree leaves are labeled numerically and these labels correspond to the comma separated list next to the names on the species tree, indicating which genes are possessed by the species, or species group, in the species tree. In this tree, we can see that the species tree indicates that the purR and rbsR are paralogs of an ancestral duplication event. Additionally, we can see that horizontal gene transfer is a likely explanation for the presence of the rpsR gene in the clade labeled *Shewanella*—because the gene is present in only one genome out of the 15 species represented by that clade, the could either be one HGT event or multiple independent deletion events.

TreeBrowser will collapse the children of a node based on a user-selectable percent identity and pick a single representative for the clade. The TreeBrowser highlights characterized homologs and selects them as representatives of the clusters.

In the species tree mode, the TreeBrowser displays a phylogenetic tree of selected gene side by side with a species tree containing all species found in the gene tree. This species tree mode allows users to manually reconcile the gene tree with the species tree in order to determine events such as, horizontal gene transfer, timing of gene duplications, orthology/paralogy relationships or differential gene loss across lineages. Clades representing one or more genes on the gene tree are labeled numerically and these labels correspond to the labels on the species tree. The names on the species tree are colored green if one or more genes present from that species are in the gene tree, red if there are no genes and orange if the species group has some members with a copy of the gene.

In the domains view, the TreeBrowser displays a view similar to the gene context view. The domain architecture of the selected gene is shown and the user can select trees based on each of these domains. This view allows users to quickly scan the members of the gene family and determine if they have shared domain architecture. Additionally, the user can view the trees associated with each of the domains of the selected gene to find evidence for gene fusions or domain shuffling.

MICROARRAY EXPRESSION PROFILING

MicrobesOnline has a wide array of tools for analyzing microarray gene expression data. The Experiment Browser allows user to search for experiments in selected genomes and/or involving specific conditions. Detailed reports on experiments are shown in the Experiment Viewer. These reports include up- and down-regulated genes, plots to assess array quality, experimental conditions and expression changes by gene functional classifications. Additionally, a Gene Expression Viewer allows users to find gene- and operon-centric views. Detailed descriptions and instructions for their use can be found on the MicrobesOnline website. One of the major uses for this microarray expression data is identifying genes which share a similar expression profile.

The gene expression profile search tool operates in two modes, a gene list view and a profile heatmap view. In both modes, the tool allows the user to select one or more genes, then calculates an average gene expression profile across the compendium of expression data and then searches for genes with a similar expression profile. In the gene list view mode, the tool returns a table of genes in descending order of their Pearson correlation coefficient to the selected gene(s) profile. The results include the correlation coefficient, gene name, gene description and buttons to either add the gene to the profile or add the gene to the analysis cart. By adding the matching gene(s) to the profile, the user can iteratively refine the profile. The profile heatmap view is a graphical heatmap representation of the selected gene(s) expression levels and those of genes identified to be similar across the microarray experiment data. The first row in the heatmap represents the input profile and genes with similar profiles are shown underneath in descending order of their Pearson correlation coefficient. The columns of the heatmap correspond to the experiments and are grouped together if the arrays are part of a series such as a time course or varying levels of a stress. The color of the cells reflects the $\log_2$ ratio value and the scale is shown to the right of the heatmap.

Results of the gene expression profile search tool can be used to identify potentially, co-regulated genes and regulatory motifs using the Motif Search tool. By adding the genes to the Gene Cart and selecting the motif search, users can potentially identify the regulatory motif.

METABOLITES AND METABOLIC PATHWAYS

Clicking on compounds in the KEGG pathway, browser now displays a MicrobesOnline internal compound page. This page typically displays the compound's image, mass, formula and structure, along with links to external databases (including KEGG) in which this compound appears. MicrobesOnline also generates a series of tables outlining the reactions in which this compound participates, delineated by whether or not the reaction occurs in the organisms under investigation. The table contains links to the reactions' EC numbers, the KEGG pathway map in which the reactions occur and the compound pages of other participants in the reactions. The compounds participating in the reactions are color-coded based on their KEGG RPAIR role (e.g. cofactors). Similarly, the gene browser now has a reactions tab that lists the pathway maps and reactions for a locus associated with the locus under investigation.

FUTURE PLANS

MicrobesOnline will continue to expand in terms of data analysis and data types supported. The next release of MicrobesOnline will include support for metagenomes, more eukaryotic genomes and an expanded set of tools for dealing with metabolites and metabolic pathways. Metagenomic analysis will feature phylogenetic trees of all genes from metagenomes and sequenced genomes and gene content-based comparisons of metagenomic samples.

FUNDING

US Department of Energy Genomics: GTL program (grant DE-AC02-05CH11231). Funding for open access charge: DOE grant.

Conflict of interest statement. None declared.

REFERENCES

1. Alm,E.J., Huang,K.H., Price,M.N., Koche,R.P., Keller,K., Dubchak,I.L. and Arkin,A.P. (2005) The MicrobesOnline Web site for comparative genomics. *Genome Res.*, **15**, 1015–1022.

2. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
3. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
4. Markowitz,V.M., Szeto,E., Palaniappan,K., Grechkin,Y., Chu,K., Chen,I.M., Dubchak,I., Anderson,I., Lykidis,A., Mavromatis,K. *et al.* (2008) The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. *Nucleic Acids Res.*, **36**, D528–D533.
5. Kazakov,A.E., Cipriano,M.J., Novichkov,P.S., Minovitsky,S., Vinogradov,D.V., Arkin,A., Mironov,A.A., Gelfand,M.S. and Dubchak,I. (2007) RegTransBase–a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. *Nucleic Acids Res.*, **35**, D407–D412.
6. Badger,J.H. and Olsen,G.J. (1999) CRITICA: coding region identification tool invoking comparative analysis. *Mol. Biol. Evol.*, **16**, 512–524.
7. Delcher,A.L., Bratke,K.A., Powers,E.C. and Salzberg,S.L. (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, **23**, 673–679.
8. Edgar R.C. PILER-CR. http://www.drive5.com/pilercr (July 1, 2009 date last accessed).
9. Bland,C., Ramsey,T.L., Sabree,F., Lowe,M., Brown,K., Kyrpides,N.C. and Hugenholtz,P. (2007) CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, **8**, 209.
10. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
11. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
12. Barrell,D., Dimmer,E., Huntley,R.P., Binns,D., O'Donovan,C. and Apweiler,R. (2009) The GOA database in 2009–an integrated Gene Ontology Annotation resource. *Nucleic Acids Res.*, **37**, D396–D403.
13. Price,M.N., Dehal,P.S. and Arkin,A.P. (2008) FastBLAST: homology relationships for millions of proteins. *PLoS ONE*, **3**, e3589.
14. Selengut,J.D., Haft,D.H., Davidsen,T., Ganapathy,A., Gwinn-Giglio,M., Nelson,W.C., Richter,A.R. and White,O. (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res.*, **35**, D260–D264.
15. Wilson,D., Madera,M., Vogel,C., Chothia,C. and Gough,J. (2007) The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res.*, **35**, D308–D313.
16. Letunic,I., Doerks,T. and Bork,P. (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res.*, **37**, D229–D232.
17. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
18. Mi,H. and Thomas,P. (2009) PANTHER Pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol. Biol.*, **563**, 123–140.
19. Nikolskaya,A.N., Arighi,C.N., Huang,H., Barker,W.C. and Wu,C.H. (2006) PIRSF family classification system for protein functional and evolutionary analysis. *Evol. Bioinform. Online*, **2**, 197–209.
20. Yeats,C., Lees,J., Reid,A., Kellam,P., Martin,N., Liu,X. and Orengo,C. (2008) Gene3D: comprehensive structural and functional annotation of genomes. *Nucleic Acids Res.*, **36**, D414–D418.
21. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
22. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
23. The UniProt Consortium. (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res.*, **37**, D169–D174.
24. Faith,J.J., Driscoll,M.E., Fusaro,V.A., Cosgrove,E.J., Hayete,B., Juhn,F.S., Schneider,S.J. and Gardner,T.S. (2008) Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.*, **36**, D866–D870.
25. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M., Marshall,K.A. *et al.* (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, **37**, D885–D890.